

LES CANCERS DU SEIN

Alexandra Aruca - Alexandre Mouton-Bistondi - Anaelle Tess Lem

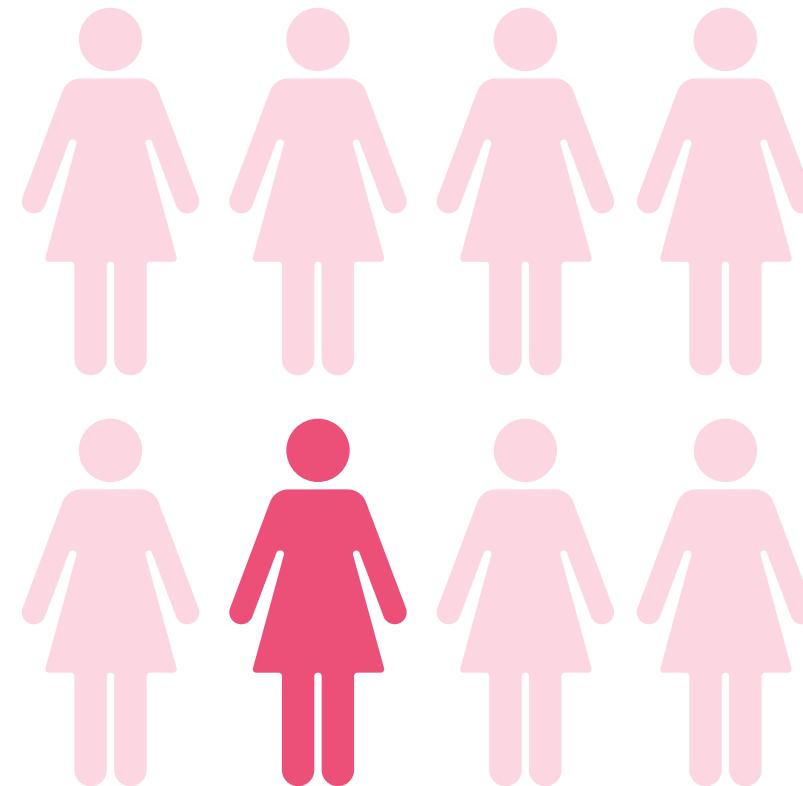
DATASET : BREAST CANCER (DIAGNOSTIC)

CONTEXTE

QUELQUES INSIGHTS



1 sur 8



2,3 millions



de femmes par an dans le monde

sensibilité de la mammographie
75-85%



jusqu'à 1 cancer sur 4 peut être manqué lors d'un examen standard

OBJECTIFS

DE NOTRE PROOF OF CONCEPT



DÉTERMINER SI UNE TUMEUR MAMMAIRE EST BÉNIGNE OU MALIGNE
à partir de variables numériques issues de prélèvements médicaux

2 OBJECTIFS

- 1 Apporter une aide au diagnostic fiable et rapide aux professionnels de santé
- 2 Tester la robustesse d'un modèle face à des données incomplètes ou bruitées, comme cela est fréquent en contexte médical

NOS DONNÉES



Source des données ?

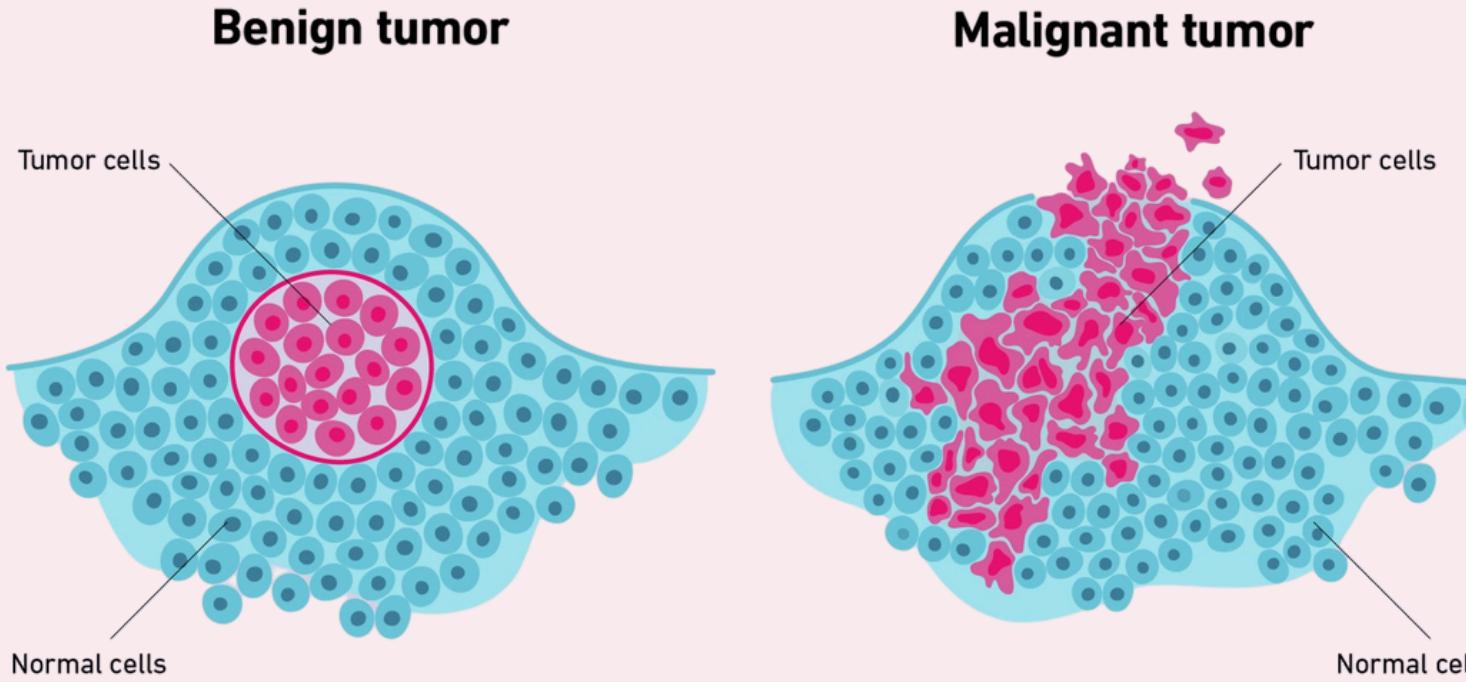
Breast Cancer Wisconsin (Diagnostic)
- *UCIrvine ML Repository*



Observations : 569 lignes = 569 patients

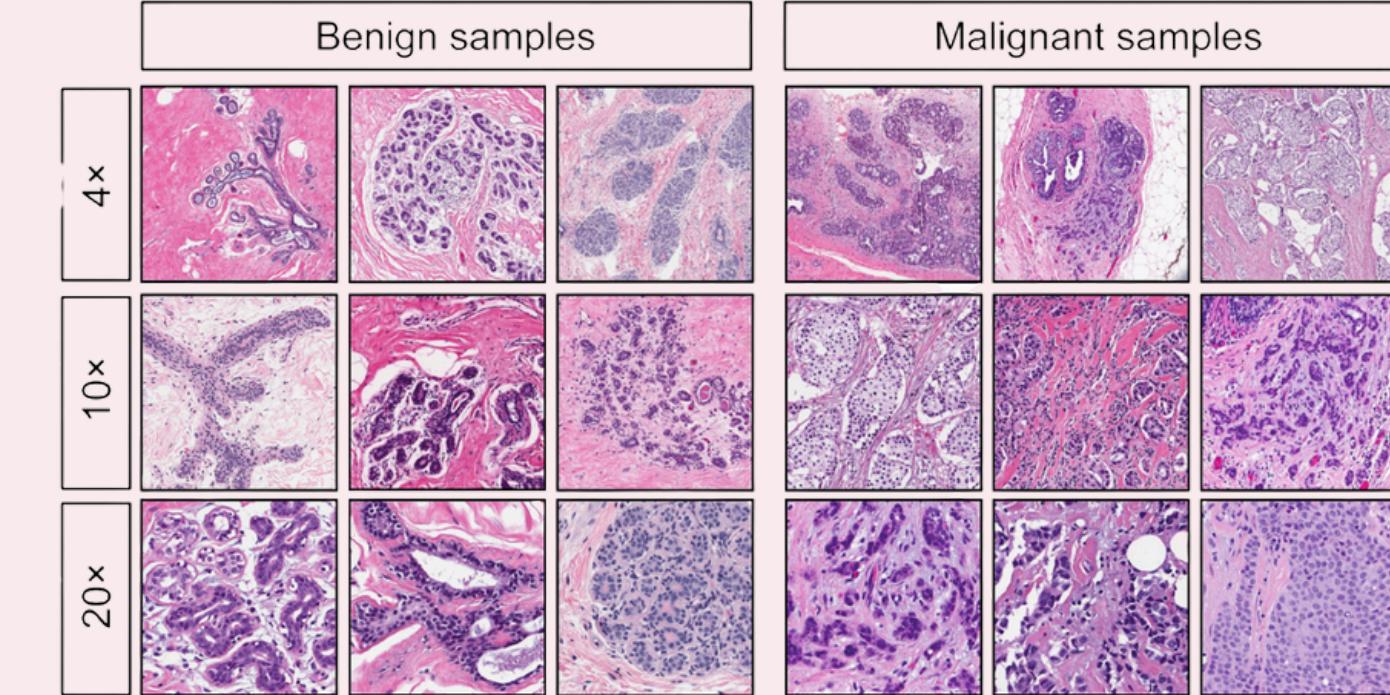
Features : 30

Type de features : données réelles 31/10/1995



COMMENT LES DONNÉES ONT-ELLES ÉTÉ COLLECTÉES ?

Images numérisées obtenues par ponction à l'aiguille fine (FNA) de tumeurs mammaires.



-> Chaque observation décrit les caractéristiques des noyaux cellulaires présents dans l'image, à travers 30 variables numériques

NOS DONNÉES



NOS DONNÉES



NOS DONNÉES

x3

id

numéro de
l'ID

diagnosis

M = maligne
B = bénigne

concavity

gravité des
creux dans le
contour

**concave
_points**

nombre de
creux dans le
contour

symmetry

symétrie des
cellules

compactness

périmètre² /
surface – 1 (densité
apparente)

perimeter

périmètre du
noyau

smoothness

variation locale des
contours

**fractal
_dimension**

complexité des
bords (fractalité)

area

surface du
noyau

radius

distance
moyenne
entre centre
et le bord

texture

écart-type
des
intensités
de l'image

- **mean** = moyenne
- **se** = erreur standard
- **worst** = valeur maximale
observée



NOS DONNÉES

x3

diagnosis

M = maligne
B = bénigne

concavity

gravité des
creux dans le
contour

**concave
_points**

nombre de
creux dans le
contour

symmetry

symétrie des
cellules

compactness

périmètre² /
surface – 1 (densité
apparente)

perimeter

périmètre du
noyau

smoothness

variation locale des
contours

**fractal
_dimension**

complexité des
bords (fractalité)

area

surface du
noyau

radius

distance
moyenne
entre centre
et le bord

texture

écart-type
des
intensités
de l'image

- **mean** = moyenne
- **se** = erreur standard
- **worst** = valeur maximale
observée



NOS DONNÉES

diagnosis

M = maligne
B = bénigne

1 cible

30 features

concavity

gravité des
creux dans le
contour

**concave
points**

nombre de
creux dans le
contour

compactness

périmètre² /
surface – 1 (densité
apparente)

perimeter

périmètre du
noyau

symmetry

symétrie des
cellules

smoothness

variation locale des
contours

fractal_
dimension

complexité des
bords (fractalité)

area

surface du
noyau

radius

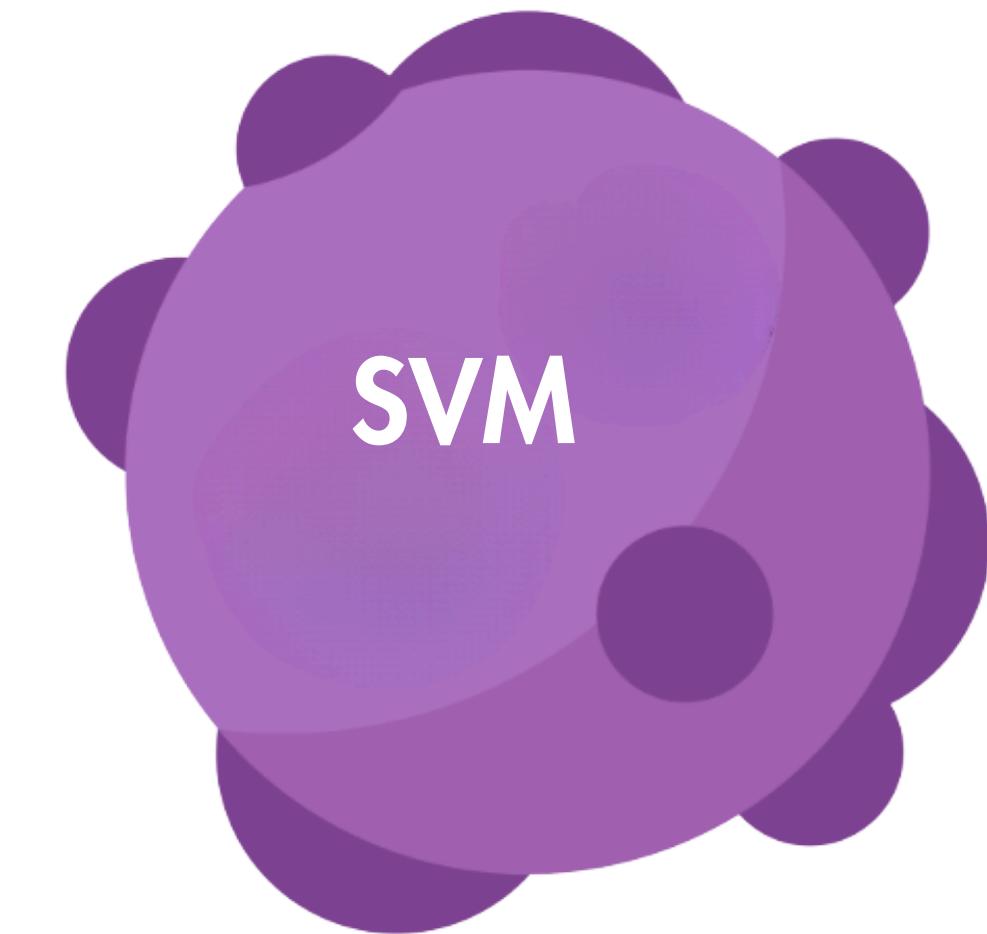
distance
moyenne
entre centre
et le bord

texture

écart-type
des
intensités
de l'image



PRÉSENTATION DES MODÈLES



PRÉSENTATION DES MODÈLES

COMMENT ÇA FONCTIONNE ?

Classifieur d'ensemble basé sur une multitude d'arbres de décision. Chaque arbre est entraîné sur un échantillon aléatoire de nos données et sur un sous-ensemble aléatoire de variables à chaque nœud. La prédiction finale est obtenue en prenant la moyenne des votes majoritaire des arbres sur la classe cible.

- PARTICULARITÉS
- RÉSULTATS



RANDOM FOREST

PRÉSENTATION DES MODÈLES

COMMENT ÇA FONCTIONNE ?

Classifieur d'ensemble basé sur une multitude d'arbres de décision. Chaque arbre est entraîné sur un **échantillon aléatoire** de nos données et sur un sous-ensemble aléatoire de variables à chaque nœud. La prédiction finale est obtenue en prenant la **moyenne des votes majoritaire** des arbres sur la **classe cible**.

▼ PARTICULARITÉS

- Peu sensible aux **variables corrélées** et au **bruit**
- Meilleure gestion des **interactions complexes** et **relations non-linéaires** entre **variables médicales**
- Fournit des **mesures d'importance** des variables

► RÉSULTATS



PRÉSENTATION DES MODÈLES



COMMENT ÇA FONCTIONNE ?

Classifieur d'ensemble basé sur une multitude d'arbres de décision. Chaque arbre est entraîné sur un **échantillon aléatoire** de nos données et sur un sous-ensemble aléatoire de variables à chaque nœud. La prédiction finale est obtenue en prenant la **moyenne des votes majoritaire** des arbres sur la **classe cible**.

- ▶ PARTICULARITÉS
- ▼ RÉSULTATS

Accuracy : 0,965
Precision : 1
Recall : 0,905
F1 Score : 0,950



PRÉSENTATION DES MODÈLES



COMMENT ÇA FONCTIONNE ?

Modèle **linéaire de classification** qui **prédit la probabilité** qu'un événement appartienne à une classe donnée en appliquant une **fonction sigmoïde** à une combinaison linéaire des variables explicatives.

- PARTICULARITÉS
- RÉSULTATS



PRÉSENTATION DES MODÈLES



COMMENT ÇA FONCTIONNE ?

Modèle **linéaire de classification** qui **prédit la probabilité** qu'un événement appartienne à une classe donnée en appliquant une **fonction sigmoïde** à une combinaison linéaire des variables explicatives.

▼ PARTICULARITÉS

- Excellente interprétabilité
- Fournit une **probabilité claire pour chaque classe**, utile pour prendre des décisions graduées (i.e seuils de diagnostic)
- Fiable sur des **données propres** et avec peu de bruit

► RÉSULTATS



PRÉSENTATION DES MODÈLES



COMMENT ÇA FONCTIONNE ?

Modèle **linéaire de classification** qui **prédit la probabilité** qu'un événement appartienne à une classe donnée en appliquant une **fonction sigmoïde** à une combinaison linéaire des variables explicatives.

► PARTICULARITÉS

▼ RÉSULTATS

Accuracy : 0,921

Precision : 0,971

Recall : 0,810

F1 Score : 0,883



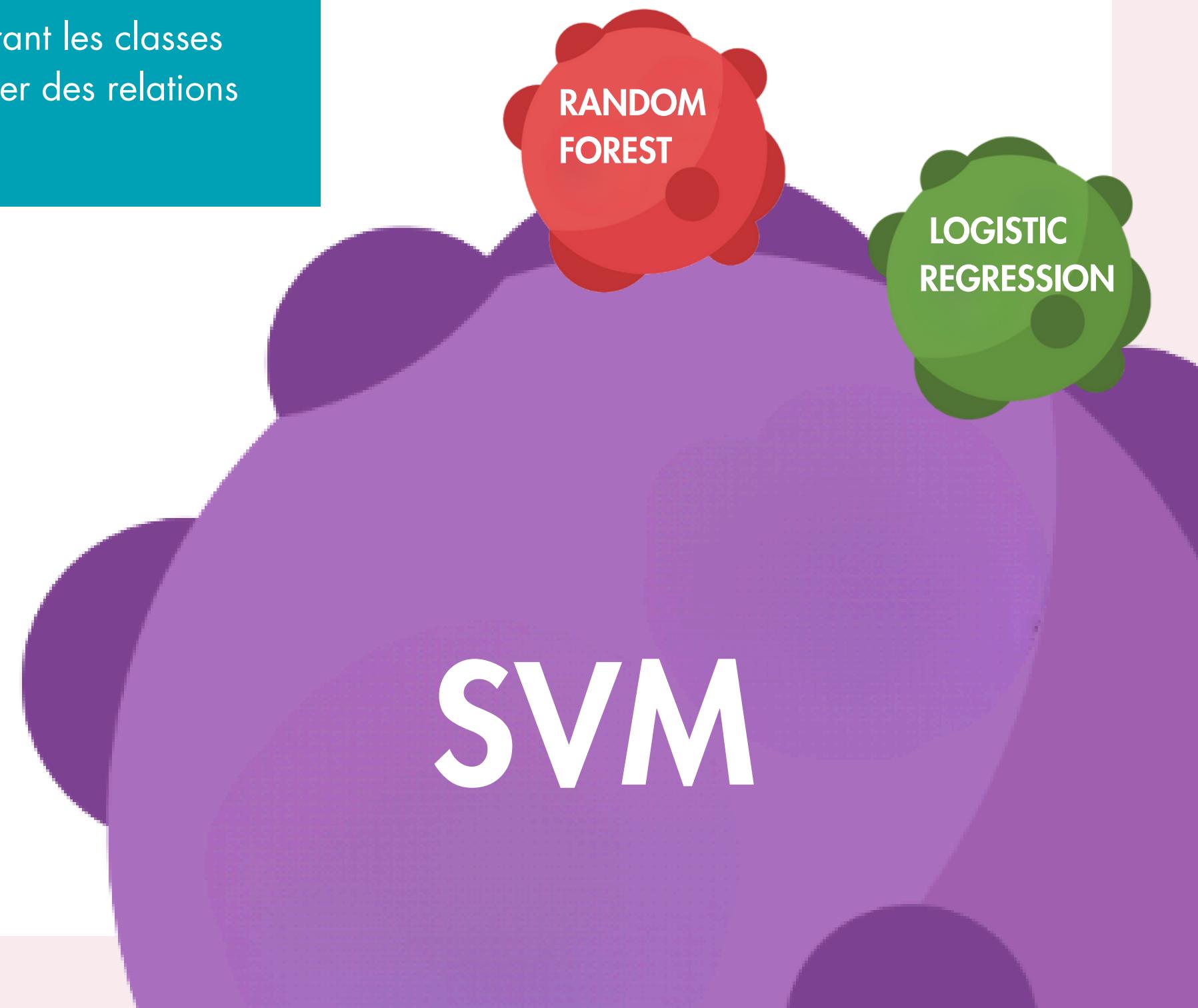
PRÉSENTATION DES MODÈLES



COMMENT ÇA FONCTIONNE ?

Modèle de **classification** qui cherche à trouver **l'hyperplan optimal** séparant les classes avec la **plus grande marge** possible. Peut utiliser des kernels pour capturer des relations **non linéaires**. Il renvoie une **probabilité** via une calibration a posteriori.

- **PARTICULARITÉS**
- **RÉSULTATS**



PRÉSENTATION DES MODÈLES



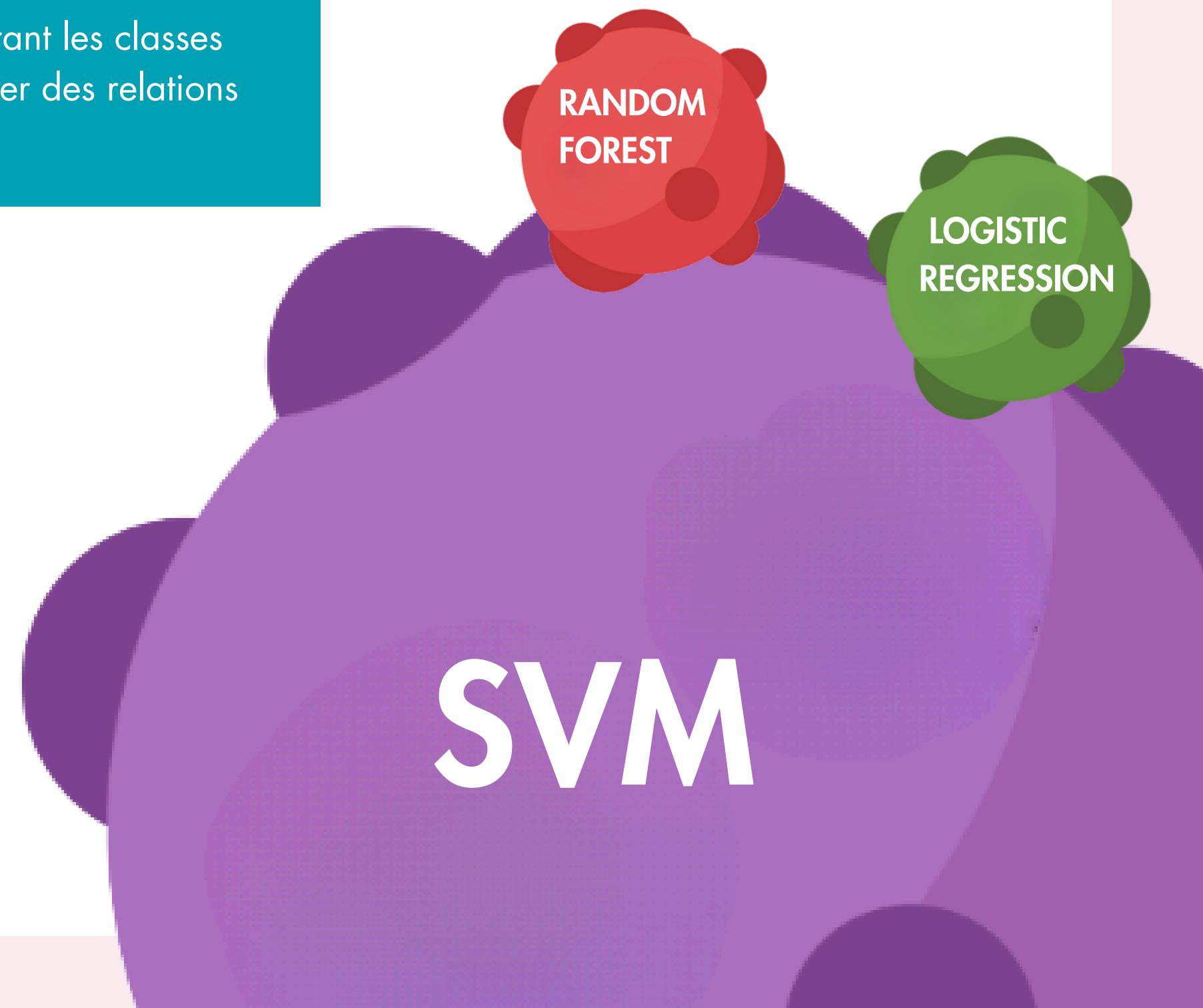
COMMENT ÇA FONCTIONNE ?

Modèle de **classification** qui cherche à trouver **l'hyperplan optimal** séparant les classes avec la **plus grande marge** possible. Peut utiliser des kernels pour capturer des relations **non linéaires**. Il renvoie une **probabilité** via une calibration a posteriori.

▼ PARTICULARITÉS

- **Résistant aux outliers**
- Très performant dans les contextes de petit volume de données

► RÉSULTATS



PRÉSENTATION DES MODÈLES



COMMENT ÇA FONCTIONNE ?

Modèle de **classification** qui cherche à trouver **l'hyperplan optimal** séparant les classes avec la **plus grande marge** possible. Peut utiliser des kernels pour capturer des relations **non linéaires**. Il renvoie une **probabilité** via une calibration a posteriori.

► PARTICULARITÉS

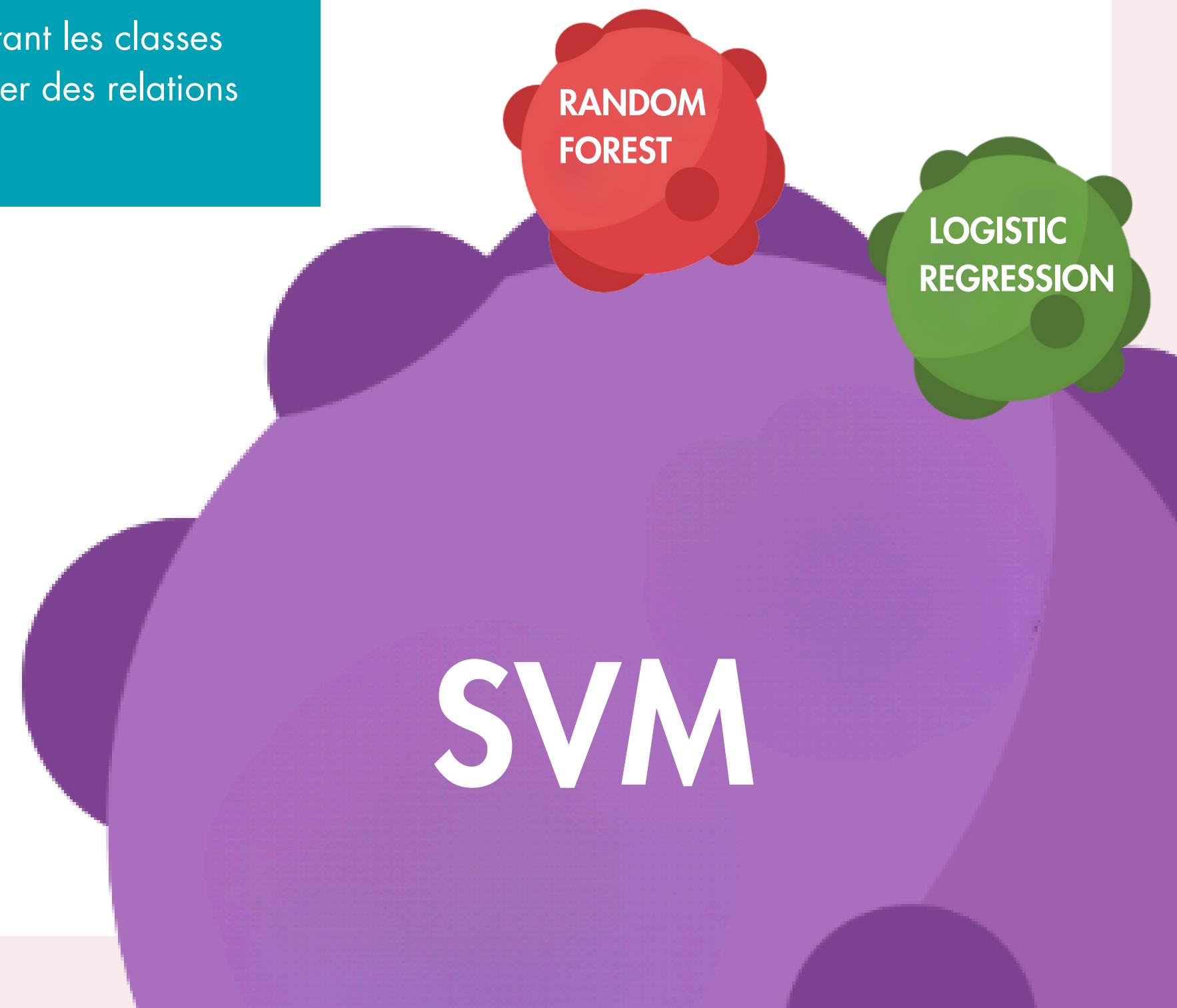
▼ RÉSULTATS

Accuracy : 0,930

Precision : 0,947

Recall : 0,857

F1 Score : 0,900



COMPARAISON DES MODÈLES



Accuracy : 0,965

Precision : 1

Recall : 0,905

F1 Score : 0,950

- Gère les NaN
- Tolère bien les données corrélées

Accuracy : 0,921

Precision : 0,971

Recall : 0,810

F1 Score : 0,883

- Requiert imputation
- Très sensible aux corrélations

Accuracy : 0,930

Precision : 0,947

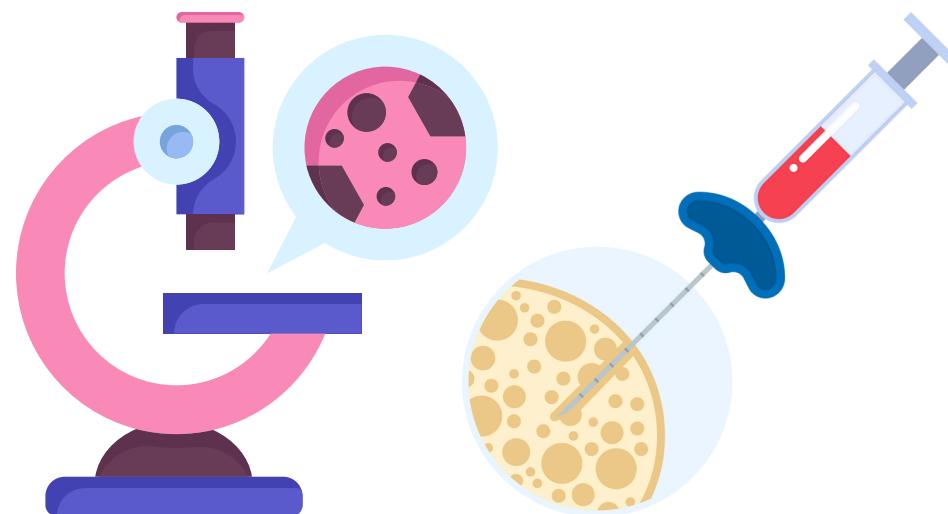
Recall : 0,857

F1 Score : 0,900

- Rejette les NaN
- Nécessite scaling et bon tuning

TESTER LA ROBUSTESSE

FACE À DES DONNÉES CLINIQUE



Lors d'une biopsie, certaines mesures peuvent être absentes :

- Échantillon trop petit
- Tissu prélevé abîmé

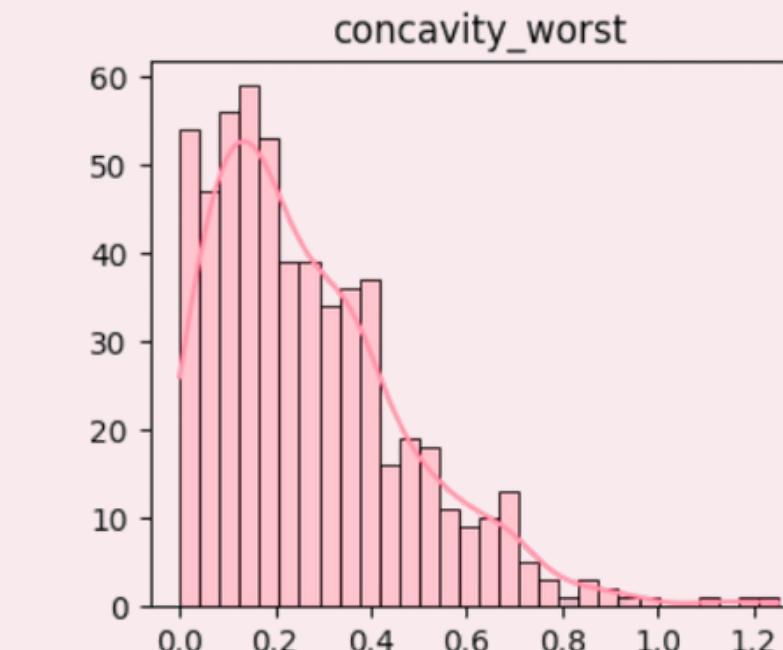
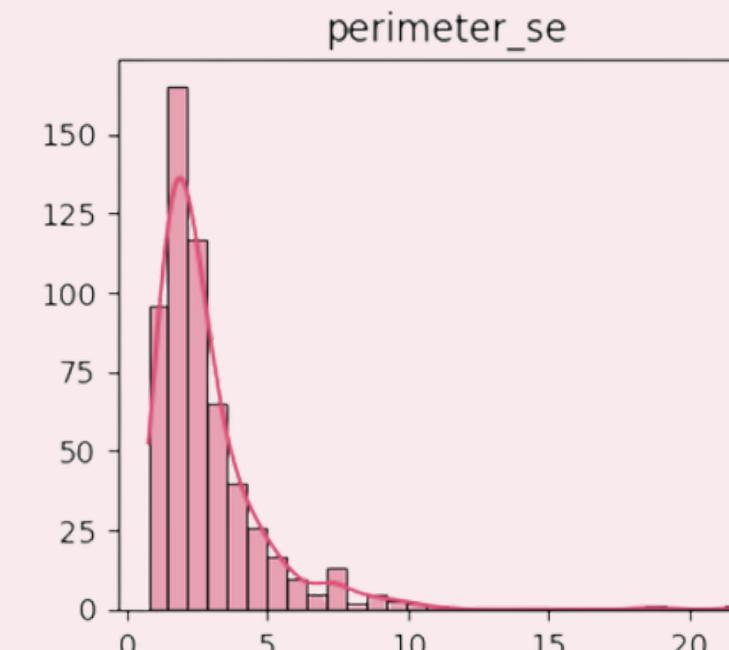
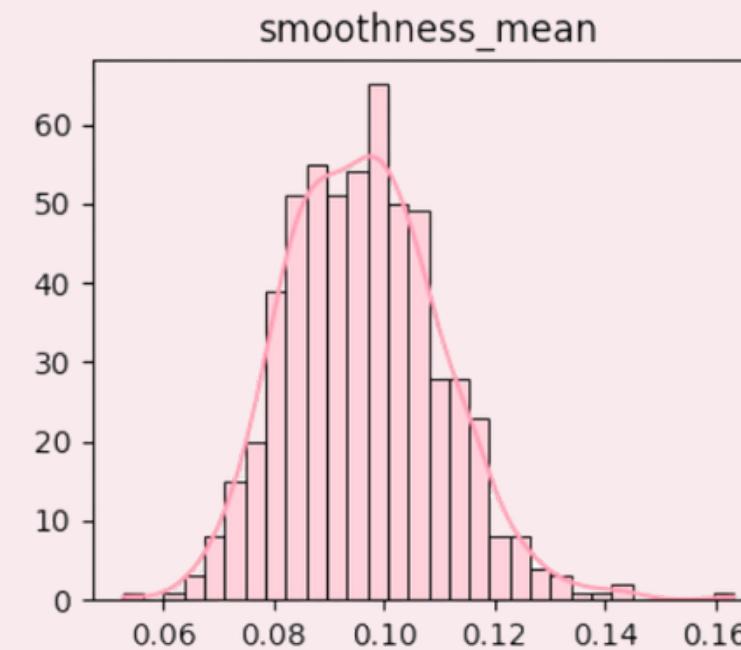
NOTRE MÉTHODE

Objectif : tester la robustesse des modèles dans ces conditions médicales réalistes

1. Création d'une fonction de simulation contrôlée pour l'ajout de NaN, avec un taux ajustable
2. **IMPUTATION** automatique des valeurs manquantes à l'aide d'une méthode cohérente
3. Évaluation des performances de nos trois modèles

IMPUTATION

FOCUS SUR LES MÉTHODES UTILISÉES



_MEAN

distributions normales (en courbe)



Imputation par la moyenne

_SE

distributions asymétriques



Imputation par la médiane

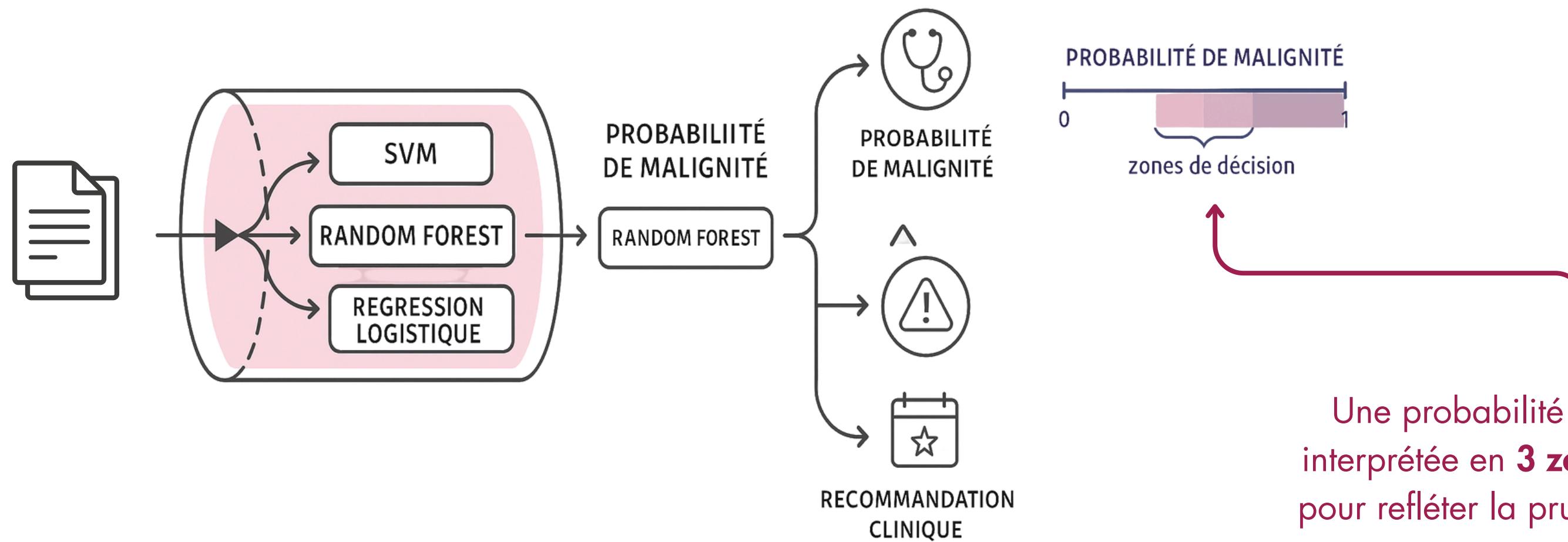
_WORST

distribution dispersées, avec de longues traines



Imputation via KNN=5

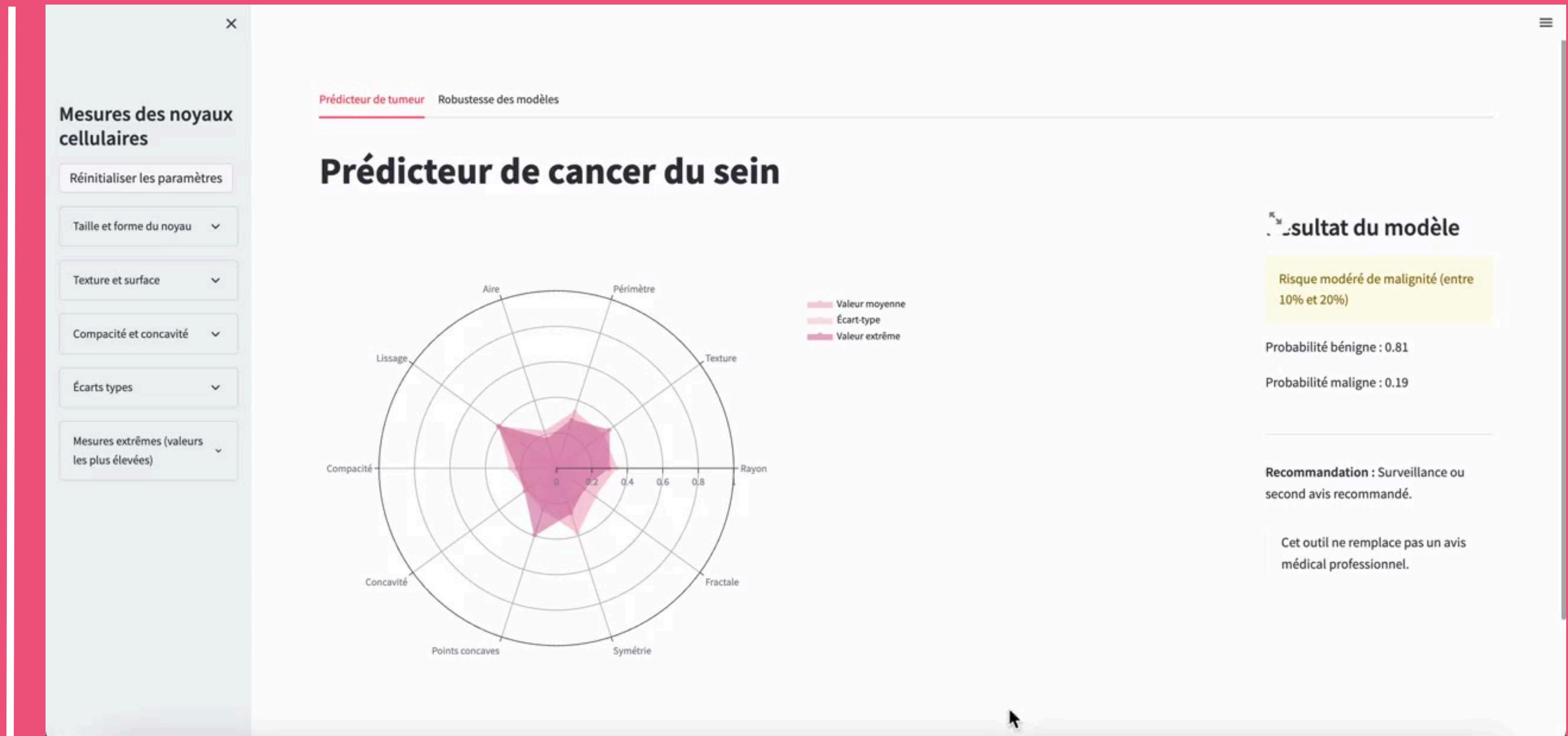
QUE FAIT NOTRE MODÈLE ?



Une probabilité (entre 0 et 1),
interprétée en **3 zones de décision**
pour refléter la prudence médicale

- > 20% : **Tumeur maligne** → Suspicion forte : IRM, intervention rapide, oncologue
- 10% < x < 20% → Risque modéré → Examen secondaire, second avis
- < 10% : **Tumeur bénigne** → Surveillance ou suivi de routine

STREAMLIT





CONCLUSION, POUR ALLER + LOIN

Rappels de notre modèle final

Modèle : Random Forest

Précision : 100% = AUCUN faux positif

→ Chaque patiente identifiée comme à risque est effectivement atteinte d'un cancer malin : très fiable pour confirmer un diagnostic

Recall : 90,5% = 1/10 cancers malins n'est pas détecté par notre modèle

→ Ce qui peut être grave si le cancer est déjà un stade avancé :



Solutions et amélioration

- Élargir le dataset avec des données réelles issues d'hôpitaux ou d'autres cohortes
- Ajouter des variables médicales clés pour re-train notre modèle dessus.
- Ajouter des variables médicales clés comme l'âge, les antécédents médicaux, certains facteurs hormonaux (cf. pilule,...)

Alexandra Aruca - Alexandre Mouton-Bistondi - Anaelle Tess Lem

MERCI POUR VOTRE ÉCOUTE



DATASET : BREAST CANCER (DIAGNOSTIC)

PRÉSENTATION DES MODÈLES



Notre dataset est considéré **petit, voire moyen** et contient uniquement des **variables numériques** et une **variable cible binaire** (diagnosis = B ou M), à **relations linéaires**. Nous avons donc un problème de **classification binaire** sur **données numériques tabulaires linéairement corrélées**, sur lesquels les modèles suivants sont particulièrement efficaces



COMPARAISON DES MODÈLES



- +
- Meilleur résultat
 - Tolère mieux les données corrélées et le bruit
 - Prédiction plus facile à interpréter pour le personnel médical

- Moins de risque de sur-apprentissage
- Attribue un poids clair à chaque variable

- Meilleur résultat
- Tolère mieux les données corrélées et le bruit
- Prédiction plus facile à interpréter pour le personnel médical

-
- Risque de sur-apprentissage
 - Moins adapté si explicabilité stricte est nécessaire

- Moins performant
- Impuissant face au bruit
- Simplification parfois trop forte des variables, ne prend pas en compte les interactions entre elles

- Très difficile à interpréter : impossible pour un médecin de justifier la décision.
- Peu intuitif à présenter
- Sensible aux paramètres de réglage

LES RÉSULTATS DE NOTRE MODÈLE AVEC BRUIT

ajout de NAN puis suppression

- **Random Forest est très robuste à la perte d'information.**
- **SVM reste performant mais devient un peu moins précis.**
- **Régression logistique est la plus sensible à la suppression de données.**

	Model	Accuracy	Precision	Recall	F1 Score
1	Random Forest	1.000	1.000	1.0	1.000
2	SVM	0.964	0.909	1.0	0.952
0	Logistic Regression	0.929	0.900	0.9	0.900

ajout de NAN puis imputation

- **Random Forest reste le modèle le plus robuste et stable, même après imputation.**
- **SVM conserve de très bonnes performances.**
- **Régression logistique est fortement affectée : elle a du mal à gérer les données imputées, surtout en termes de rappel (risque élevé de faux négatifs).**

	Model	Accuracy	Precision	Recall	F1 Score
1	Random Forest	0.974	1.000	0.929	0.963
2	SVM	0.965	0.975	0.929	0.951
0	Logistic Regression	0.816	0.818	0.643	0.720